Multiple Event Audio Tagging

Debojyoti Misra
IIT Kanpur
Electrical Engineering Department
Roll No. : 22104402
debojyotim22@iitk.ac.in

1 Introduction

The task given to me is a multiple event audio detection. Audio being a sequence of data is not that easy for Deep Neural Networks(DNN) to classify. To overcome this problem the first thing that comes to mind is a Recurrent Neural Network(RNN). But in case of longer sequences of data RNN's face problems of short term memory and vanishing gradient. These problems are addressed with two variatons of RNN's, Long Short Time Memory(LSTM) Networks and Gated Recurrent Units(GRU).In this problem of polyphonic detection I use a Bidirectional LSTM(Bi-LSTM) Network for embedding followed by a DNN for multi-label classification.

2 Literature Survey

Parascandolo Et al. in his work[3] first used Bi-LSTM Networks in the task of polyphonic detection. In this work they used a Bi-LSTM Network to embed a Mel Spectrogram which is then passed into a DNN to classify different classes present in each timestep. Later Emre Et al. used Convolutional Recurrent Neural Networks(CRNN) for polyphonic detection in his work[1]. They used a Convolutional Neural Network(CNN) to generate features from Mel Spectograms which was again passed into an RNN to create embeddings. Mortiz Et al. in his work[2] used a Time Delay Neural Network(TDNN) to create embeddings.

3 My Method

I tried to adapt [3] to solve this task. The data I was given is 10000 Mel Spectrograms having 64 frequency components over 1000 time-frames.First I Transposed the data sample. I converted each data sample into 20 equal parts along the time axis of 50 time-frames each and stacked them together. The dimension that we get after this for each sample is (20,50,64). This is done because I am passing it to a Bi-LSTM Network.This (20,50,64) shaped sample is then Normalised along the second axis. Performance of LSTM Networks get significantly better if we pass smaller chunks to them. The Bi-LSTM network gives us (50,512) embeddings every chunk of a data sample. We pass it to a Time Distributed Neural Network which generates pobability of each class being present for every time-step in a chunk

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

of the data sample. For each chunk our output has a shape of (50,11) as we have 11 classes to identify.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional_2 (Bidirectio  (None, 50, 512)          657408
 nal)

 bidirectional_3 (Bidirectio  (None, 50, 512)          1574912
 nal)

 time_distributed_3 (TimeDis  (None, 50, 256)          131328
 tributed)

 time_distributed_4 (TimeDis  (None, 50, 64)           16448
 tributed)

 time_distributed_5 (TimeDis  (None, 50, 11)           715
 tributed)

=================================================================
Total params: 2,380,811
Trainable params: 2,380,811
Non-trainable params: 0
_________________________________________________________________
```

The Ground Truth given to us had 11 classes for each 1000 time-frame which is also Transposed and broken in chunks of 50 time-frames each and stacked together to create an array of (20,50,11) in which the presence of each class is represented by 1 in every time-frame and as we have 11 classes,each time-frame is a multi-hot vector of 11 dimensions. We used Binary Cross Entropy Loss as the Objective. This Objective was optimized with the help of an Adam Optimizer with learning rate of 0.0001.

4 Evaluation Matrics(On test data samples)

Class	Precision	Recall	F1 Score	Accuracy
Alarm Bell Ringing	0.46	0.5	0.48	0.92
Blender	0.45	0.5	0.47	0.9
Cat	0.48	0.5	0.49	0.96
Dishes	0.49	0.5	0.49	0.97
Dog	0.48	0.5	0.49	0.96
Electric Shaver Toothbrush	0.46	0.5	0.48	0.92
Frying	0.43	0.5	0.46	0.86
Running Water	0.45	0.5	0.47	0.9
Silence	0.82	0.9	0.84	0.88
Speech	0.8	0.72	0.72	0.76
Vaccum Cleaner	0.45	0.5	0.47	0.9

Average Precision: 0.52
Average Recall: 0.56
Average F1 Score: 0.53

Average Accuracy: 0.9

5 Observation and Discussion

In these audio files we observe that the proportion of Speech and Silence is much more than that of the other 9 labels which is very practical.This results in this method giving far better results for the former 2 classes than the later 9.

References

[1] Emre Çakır et al. "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (Feb. 2017). DOI: 10.1109/TASLP.2017.2690575.

[2] Niko Moritz et al. "Acoustic scene classification using time-delay neural networks and amplitude modulation filer bank features". In: Jan. 2016.

[3] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. "Recurrent neural networks for polyphonic sound event detection in real life recordings". In: Mar. 2016, pp. 6440–6444. DOI: 10.1109/ICASSP.2016.7472917.